Language Resource Addition: Dictionary or Corpus?

Shinsuke Mori Graham Neubig

Kyoto University NAIST

2014 May 29

Table of Contents

NLP for Applications

- Machine learning approach
  1. Annotation standard
  2. Language resource (Texts with annotations)
  3. Classifiers
- High accuracy in the general domain
  - We have enough large annotated data
- Not sufficiently accurate for various texts
  - Achieve a high accuracy *by all means!!*

Language Resource Addition for ML-based NLP

Language resource addition never betrays!!

- As dictionary entries
  - Without context $\Rightarrow$ Improve NLP
  - Easy for tool users $\because$ You just edit the dictionary.
- As an annotated corpus
  - Not easy for tool users $\because$ You need re-training.
  - With context $\Rightarrow$ Improve more?

Task for Experiments

- Japanese morphological analysis = WS + PT

Word segmentation (WS)

ex.) 吾輩は猫である　　　　I am a cat

$\Downarrow$

吾輩 は 猫 で あ る

Part-of-speech tagging (PT)

ex.) 吾輩 は 猫 で あ る

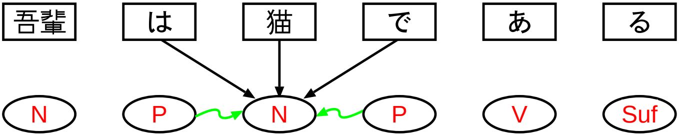
$\Downarrow$

N P N P V Suf

- Most ambiguity lies in WS

Sequence-based Approach (SB)

- MeCab: CRF-based joint method [Kudo 04]



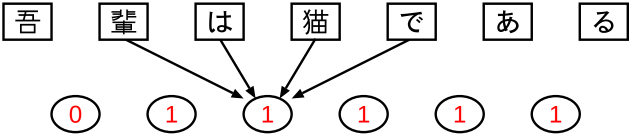
- refers to the word to be tagged $\boldsymbol{w}$, the word sequences to its left $\boldsymbol{w}_-$ and right $\boldsymbol{w}_+$, and their POS
- requires fully annotated language resources

ex.) 吾輩/**N** は/**P** 猫/**N** で/**P** あ/**V** る/**Suf**

Cf. [Tsuboi 08]

Pointwise Approach (PW)

- KyTea: 2-step pointwise method (SVM or other) [Neubig 11]
  - Word segmentation ⇒ POS tagging



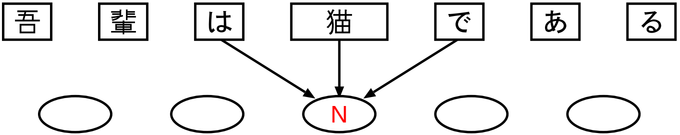
- refers to only the word to be tagged $\boldsymbol{w}$, and the character sequences to its left $\boldsymbol{c}_{-}$ and right $\boldsymbol{c}_{+}$
- never refers to any estimated values!
- is trainable from partially annotated language resources

ex.) 吾輩は 猫 である

no annot. no annot.

Pointwise Approach (PW)

- KyTea: 2-step pointwise method (SVM or other) [Neubig 11]
  - Word segmentation ⇒ POS tagging



- refers to only the word to be tagged $\boldsymbol{w}$, and the character sequences to its left $\boldsymbol{c}_{-}$ and right $\boldsymbol{c}_{+}$
- never refers to any estimated values!
- is trainable from partially annotated language resources

$\boxed{\text{ex.) 吾輩は 猫/N である}}$

no annot. no annot.

Dictionary or Corpus

Dictionary

word1/POS1,POS2
word2/POS2,POS3
⋮

Corpus

left context	word1/POS1	right context
left context	word1/POS2	right context
left context	word2/POS2	right context
left context	word2/POS3	right context
	⋮	

- Unknown words are found in real texts with contexts

Experimental Setting

1. BCCWJ (Balanced Corpus of Contemporary Written Japanese) [Maekawa 08]

Corpus		
Domain	#words	
General	784k	(Core Data - Yahoo!QA)
General + Web	898k	(Core Data)
Web for test	13.0k	
Dictionary		
Domain	#words	Coverage (word/POS)
General	29.7k	96.3%
General + Web	32.5k	97.9%

MA and method

- Morphological analyzer
  1. MeCab: CRF-based joint method [Kudo 04]
  2. KyTea: 2-step pointwise method [Neubig 11]
- Adaptation strategies
  1. No adaptation: Use the corpus and the dictionary in the general domain.
  2. Dictionary addition (no re-training): Add words appearing in the Web training corpus to the dictionary (MeCab only).
  3. Dictionary addition (re-training): + estimate the weights on the general domain training data.
  4. Corpus addition: Add annotated sentences in the Web training corpus and train the parameters.

Accuracy Mesurement

- N_{REF}: the number of word-POS pairs in the correct sentence
- N_{SYS}: in the system output
- N_{LCS}: the length of the LCS (longuest common subsequence)

$$\text{Recall} = \frac{N_{LCS}}{N_{REF}}, \quad \text{Prec.} = \frac{N_{LCS}}{N_{SYS}}.$$

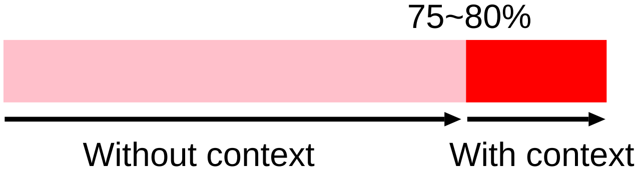
- F-measure: the harmonic mean of the Recall and the Prec.

$$F = \left\{ \frac{1}{2}(R^{-1} + P^{-1}) \right\}^{-1} = \frac{2N_{LCS}}{N_{REF} + N_{SYS}}.$$

Word Segmentation Accuracy

Adaptation strategy	MeCab	KyTea
No adaptation	95.20%	95.54%
Dict. addition (no re-training)	96.59%	-
Dict. addition (re-training)	96.55%	96.75%
Corpus addition	96.85%	97.15%

- Dictionary addition: +1.35% (MeCab), +1.21% (KyTea)
- Corpus addition: +0.30% (MeCab), +0.40% (KyTea)



Realistic Cases

- The previous experiments are somewhat artificial or *in-vitro*
  - Full annotation required

    ex.) 吾輩/**N** は/**P** 猫/**N** で/**P** あ/**V** る/**Suf**

- Two real adaptation scenarios or *in-vivo*
  - Partial annotation

    ex.) 吾輩は 猫/**N** である

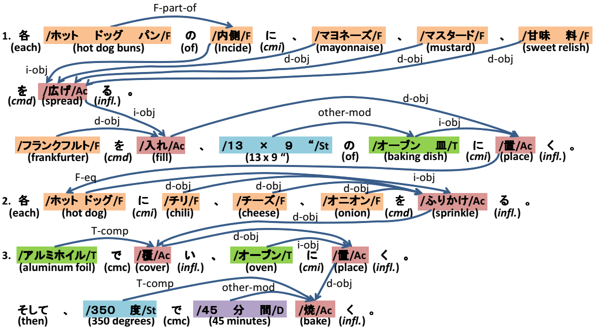
    no annot. no annot.

  - Only KyTea (MeCab does not support such data)
- focusing on word segmentation where most ambiguity lies

Case 1: Recipe Text Analysis

for Procedural Text Understanding

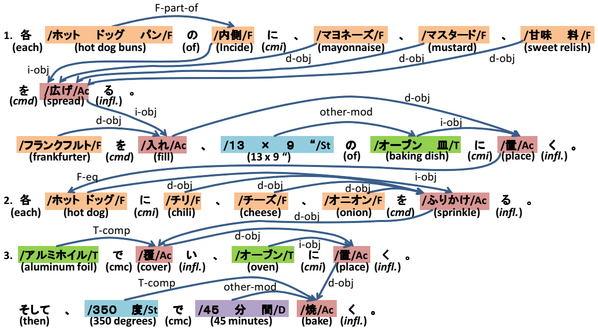
- Recipe flow graph corpus [Mori 14] (05/29 Session: P34 - Corpora and Annotation)



- Specifications

	#Sent.	#NEs	#Words	#Char.
Training	1,760	13,197	33,088	50,002
Test	724	–	13,147	19,975

Recipe flow graph corpus



- "Meaning representation" of cooking instructions
- Important terms for cooking (recipe NEs) are annotated with types and correctly segmented into words

Three Adaptation Methods (1/2)

1. No Adapation
2. Dictionary: Use the training data as a dictionary.
   2.1 Extract recipe NEs from the training data,
   ex.) /ホット ドッグ/F, /チリ/F, /チーズ/F,
   /オニオン/F, /ふりかけ/Ac,
   /ホット ドッグ/F, /アルミ ホイル/F, /覆/Ac
   2.2 Make a dictionary containing the words in these NEs,
   ex.) ホット, ドッグ, チリ, チーズ, オニオン,
   ふりかけ, アルミ, ホイル, 覆
   2.3 Use the dictionary as the additional language resource to train the model.

Three Adaptation Methods (2/2)

3. Corpus: Use the training data as partial annotation

1. Extract n occurrences at maximum of the recipe NEs

各	/ホット ドッグ/F	に	チリ	、	チーズ	、
(each)	(hot dog)	(*cmi*)	(chili)	,	cheese	,

オニオン	を	ふりかけ	る
onion	(*cmd*)	(sprinkle)	(*infl.*)

/ホット ドッグ/F	を	アルミ ホイル	で	覆	う
(hot dog)	(*cmd*)	(aluminum foil)	(by)	(cover)	(*infl.*)

2. Convert them into partially segmented sentences
   - both edges and the inside of the NEs are annotated with word boundary information.

ex.) 各|ホ-ッ-ト|ド-ッ-グ|に␣チ␣リ␣、⋯,
|ホ-ッ-ト|ド-ッ-グ|を␣ア␣ル␣ミ⋯,

   - "|": boundary, "-": not boundary, "␣": no information

3. Train the model with this partially annotated data

Word Segmentation Accuracy

Strategy	#occurrences max.(n)	#occurrences average	#words	WS F-measure BCCWJ	WS F-measure Recipe
No adaptation	–	–	0	98.87%	94.35%
Dictionary	–	–	1,999	98.90%	94.54%
Corpus (partial annotation)	1	1.00	1,999	98.89%	95.56%
	2	1.60	3,191	98.89%	95.81%
	3	2.02	4,046	98.89%	95.94%
	4	2.36	4,727	98.89%	96.01%
	8	3.26	6,523	98.89%	96.07%
	16	4.26	8,512	98.89%	96.14%
	32	5.10	10,203	98.89%	96.21%
	64	5.77	11,542	98.89%	96.28%
	∞	6.60	13,197	98.89%	96.29%

- Partial annotation is better than dictionary addition
- The degree of improvement shrinks as n increases.

Case 2: Patent Text Analysis
for Machine Translation

KWIC ⇒ Distributional Analysis ⇒ Partial annotation

"嵌 合" (Freq = 49)

1. こ␣れ␣ら|嵌-合|用␣ロ␣ッ␣ク
2. 7␣c␣が|嵌-合|方␣向␣に␣向
3. 自␣在␣に|嵌-合|す␣る␣回␣転

1. Extract unknown word candidates based on the distributional similarity from a raw corpus in the target domain [Mori 96]

2. Sort them in the descending order of the expected frequencies

3. Annotate three occurrences with word boundary information

※ In the beginning, $\mathbf{4} \leq n \leq \mathbf{8}$ in the case 1 result

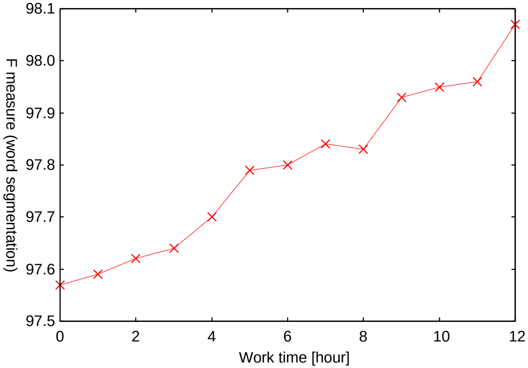
Invention Disclosure Corpus

- Corpus specifications

	#Sent.	#Words	#Char.
Raw	31,862	–	2,018,082
Test	500	20,658	32,139

1. One hour annotation
2. Word segmentation model estimation
3. Accuracy measurement
4. Goto 1.

Result



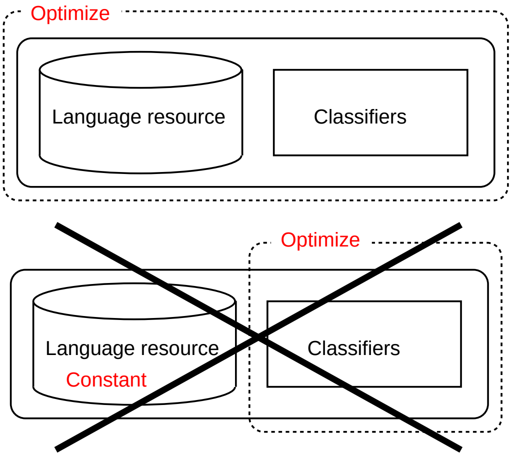
- The accuracy gets higher as we add partial annotations.
- 12 hours of annotation eliminated 20% of the errors.
- The final F-measure is as high as the general domain.
- We can improve more by only more annotator's work.

Conclusion

- Corpus > Dictionary
  - Context information
  - Three occurences in vivo
- Never throw away the context when you find an unknown word
- NLP trainable from partial annotations
  - Allows to focus on unknown (or important) words
  - Must be as accurate as the state-of-the-art NLP

Take Home Message

- Optimize the entire process with a flexible analyzer



References

- Kudo, T., Yamamoto, K., and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 230–237 (2004)
- Maekawa, K.: Balanced Corpus of Contemporary Written Japanese, in *Proceedings of the 6th Workshop on Asian Language Resources*, pp. 101–102 (2008)
- Mori, S. and Nagao, M.: Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis, in *Proceedings of the 16th International Conference on Computational Linguistics* (1996)

- Mori, S., Maeta, H., Yamakata, Y., and Sasada, T.: Flow Graph Corpus from Recipe Texts, in *Proceedings of the Nineth International Conference on Language Resources and Evaluation* (2014)
- Neubig, G., Nakata, Y., and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (2011)
- Tsuboi, Y., Kashima, H., Mori, S., Oda, H., and Matsumoto, Y.: Training Conditional Random Fields Using Incomplete Annotations, in *Proceedings of the 22th International Conference on Computational Linguistics* (2008)